

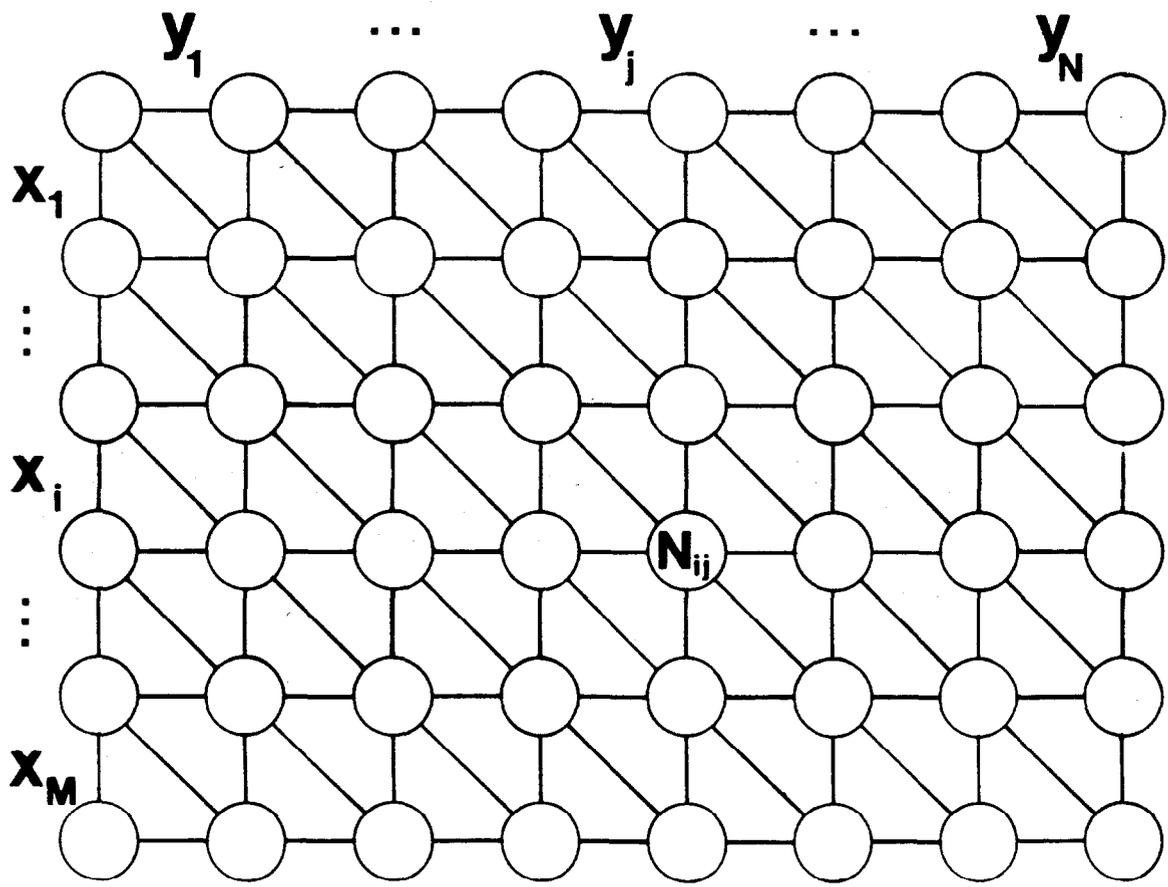
**Definitions.** What is one trying to find or optimize?

**Algorithms.** Can one find the proposed object optimally or in reasonable time?

**Statistics.** Can the result be explained purely by chance?

Generally there is a tension between biologically reasonable definitions of a problem and tractable algorithms and statistics. Balancing these considerations requires judgement.

**Applications.** Can one make biologically relevant discoveries?

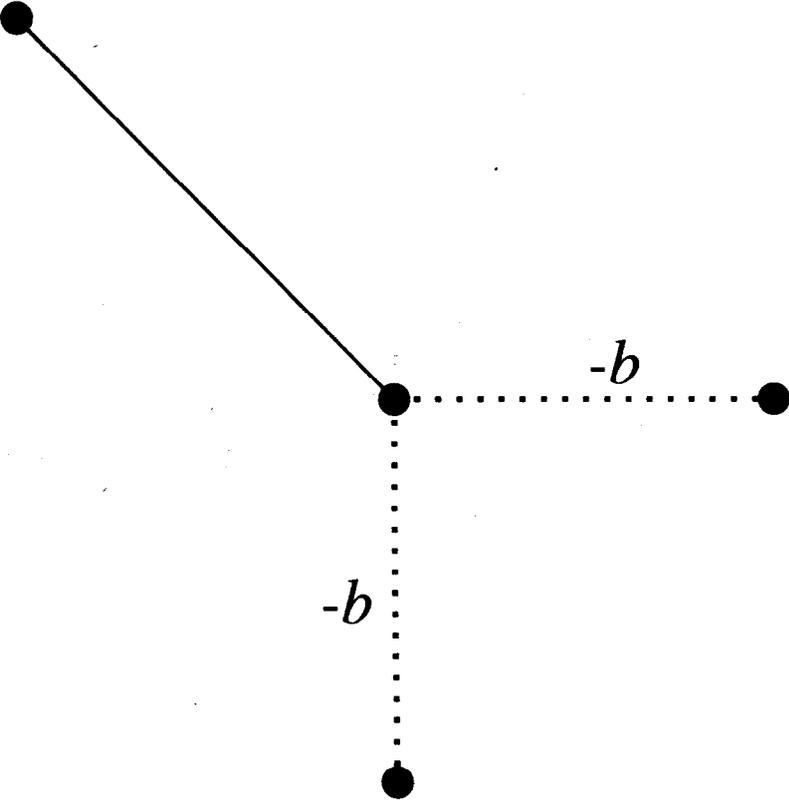


## The BLOSUM-62 Matrix

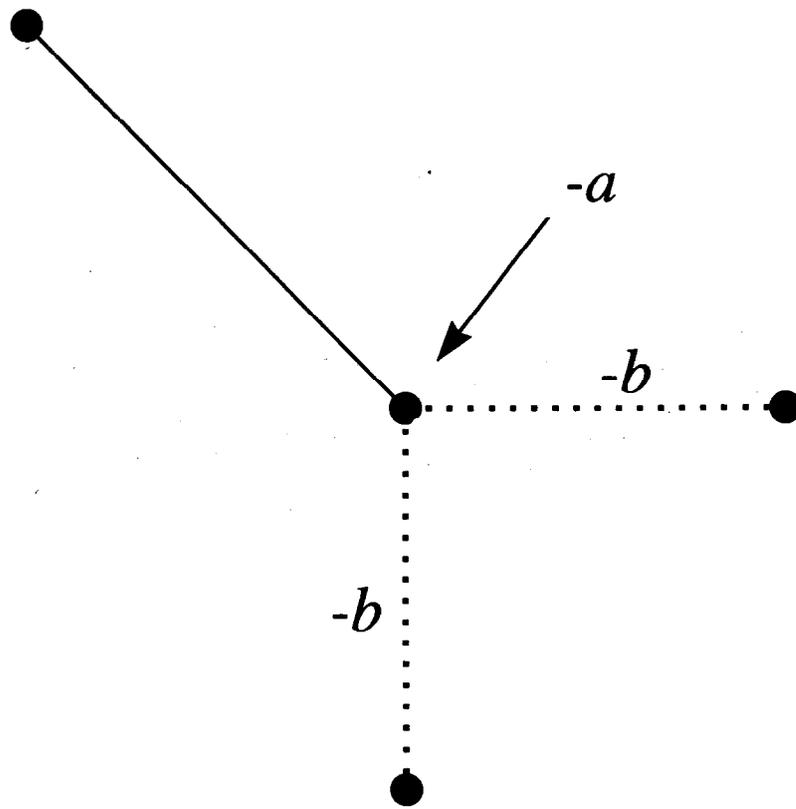
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

(PNAS 89:10915-10919)

# Length-Proportional Gap Costs



# Affine Gap Costs



# THE STATISTICAL THEORY

The various amino acids, indexed by letters  $i$  or  $j$ , occur randomly and independently with respective probabilities

$$p_1, p_2, \dots, p_i, \dots, p_{20}.$$

The *score* for aligning amino acids  $i$  and  $j$  is  $s_{ij}$ . A substitution score matrix is then made up of the scores

$$s_{1,1}, s_{1,2}, \dots, s_{ij}, \dots, s_{20,20}.$$

## NEGATIVE EXPECTED SCORE

Score matrices used to seek local alignments of variable length should have a negative expected score:

$$\sum_{i,j} p_i p_j s_{ij} < 0.$$

Otherwise, alignments representing true homologies will tend to be extended with biologically meaningless noise.

## LOG-ODDS SCORES

The scores of *any* scoring matrix can be written in the form

$$s_{ij} = \left[ \ln \frac{q_{ij}}{p_i p_j} \right] / \lambda = \log \frac{q_{ij}}{p_i p_j}$$

where the  $q_{ij}$  are a set of *target frequencies* for aligned amino acid pairs (PNAS 87:2264-2268).

## NORMALIZED SCORES

Different scoring systems may be compared by means of normalized scores (J. Mol. Evol. 36:290-300). If  $S$  is the score for a local alignment, the *normalized* score (expressed in bits) may be defined as

$$S' = (\lambda S - \ln K) / \ln 2$$

where  $\lambda$  and  $K$  are parameters associated with the local alignment scoring system. For alignments without gaps,  $\lambda$  and  $K$  may be calculated (PNAS 87:2264-2268); otherwise, they must be estimated (Meth. Enzymol. 266:460-480). The expected number of distinct local alignments with normalized score at least  $S'$  bits is  $N2^{-S'}$ , where  $N$  is the size of the search space.

## STATISTICAL SIGNIFICANCE

A normalized score  $S'$  is statistically significant (with E-value  $E$ ) if it exceeds

$$\log N/E$$

where  $N$  is the size of the search space (PNAS 87:2264-2268). For a typical current database search,  $N \approx 10^{10} \approx 2^{33}$ , so a normalized score of about 38 bits is statistically significant.

## SEARCH SPACE SIZE

Database length  $\approx 30,000,000 \approx 2^{25}$

Query  
length  
 $\approx 2^8$

$$N \approx 2^{33}$$

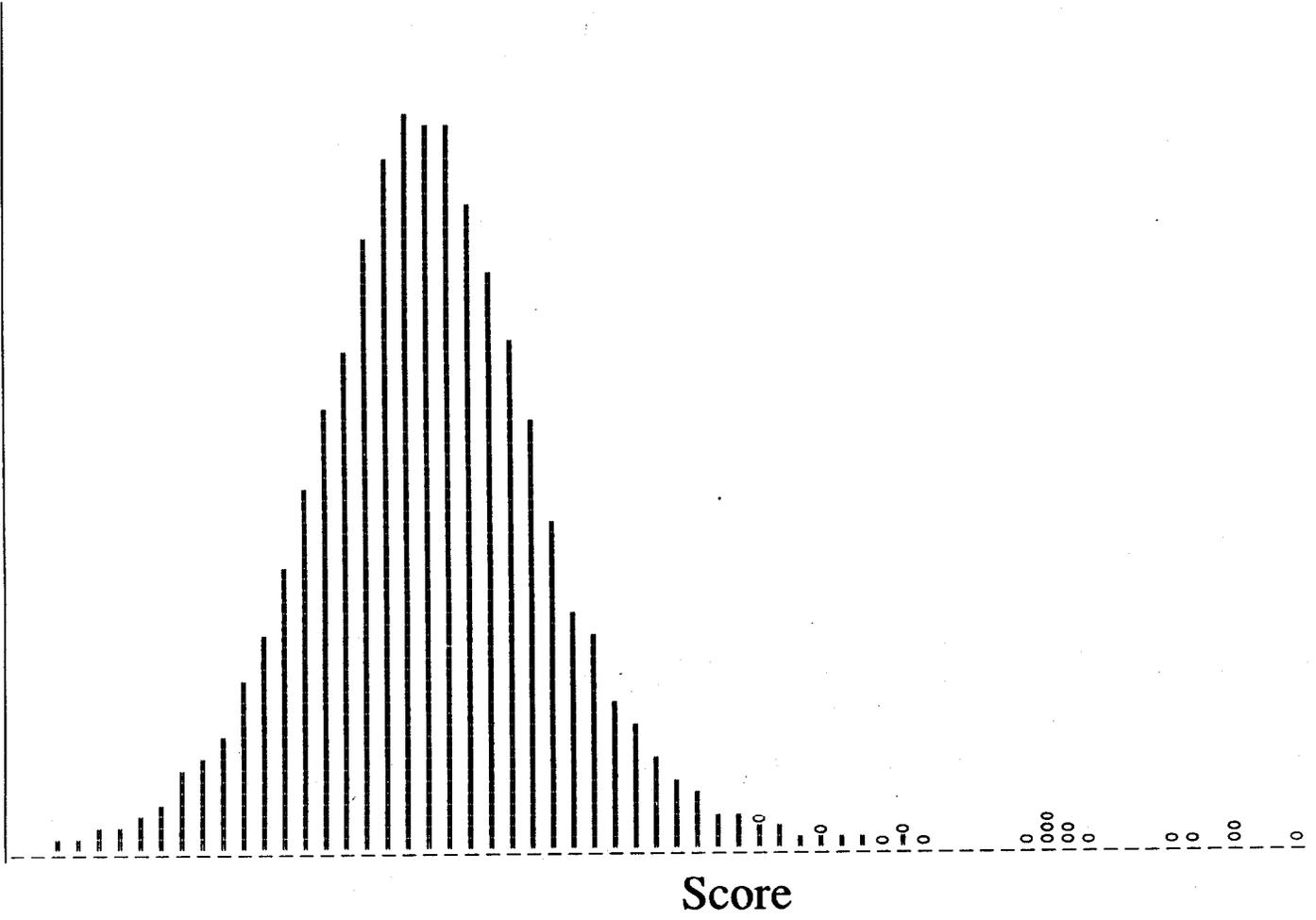
## SOME TYPICAL NUMBERS

Protein length	250	$\approx 2^8$
Database length	30,000,000	$\approx 2^{25}$

High random HSP scores:

Database search	33 bits	( $E \approx 1.00$ )
	39 bits	( $E \approx 0.02$ )
Pairwise comparison	16 bits	( $E \approx 1.00$ )
	22 bits	( $E \approx 0.02$ )

#



\*

Human beta-globin VHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMG  
LTPEE VT LWGKVVN VGGGEALGRLLVVYPWTQRFFESFGDLS PDA MG  
Ring-tailed lemur beta-globin TFLTPEENGHVTSLWGKVVNVEKVGGEALGRLLVVYPWTQRFFESFGDLSSPDAIMG

NPKVKAHGKKVLGAFSDGLAHLNLDKGTFFATLSELHCDKLVHDPENFRLLGNVLVLCVLAHFGKEFTPPVQAAAYQKVVAGVANALAHKYH  
NPKVKAHGKKVL AFS GL HLDNLDKGTFA LSELHC LVHDPENF LLGNVLV VLAHFG F P QAA QKVV GVANALAHKYH  
NPKVKAHGKKVLSAFSEGLHLDNLDKGTFAQLSELHCVLHDPENFKLLGNVLVIVLAHFGNDFSPQTQAAFQKVVTVGVANALAHKYH

Human beta-globin VHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN  
V T E SA LWGK N DE G AL R L VYPWTQR F FG LS P A MGN  
Goldfish beta-globin VEWTDAERSAIGLWGKLNPDDELGPQALARCLIVYPWTQRVYFATFGNLSSPAAIMGN

PKVKAHGKKVLGAFSDGLAHLNLDKGTFFATLSELHCDKLVHDPENFRLLGNVLVLCVLAHFG-KEFTPPVQAAAYQKVVAGVANALAHKYH  
PKV AHG V G DN K T A LS H KLHVPD NFRLL A FG F VQ A QK V AL YH  
PKVAAHGRTVMGGLERAIKNMDNIKATYAPLSVMHSEKLVHDPDNFRLLADCITVCAAMKFGP SGFNADVQEAQKFLSVVVSALCRQYH

Human beta-globin VHLTPEEKSAVTALW----GKVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVK  
L V W G N VGE L F F S P V  
Bloodworm globin IV MGLSAAQRQVVASTWKDIAGSDNGAGVGKECF TKFLSAHHDIAAVF-GFSGAS-----DPGVA

AHGKKVLGAFSDGLAHL-DNLKGTFFATLSELHCDK----LVHDPENFRLLGNVLVLCVLAHFGKEFTPPVQAAAYQKVVAGVANALAHKYH  
G KVL D HL D K K H E F LG L H G T A A AL  
DLGAKVLAQIGVAVSHLGDEGKMAEMKAVGVRHKGYGYKHIKAEYFELGASLLSAMEHRIGGKMTAAAKDAWAAAYADISGALISGLQ

Human beta-globin VHLTPEEKSAVTALWG--KVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP  
V T V KN L P F P NP  
Soybean leghemoglobin VAFTEKQDALVSSSFEAFKANIPQYSVVVYFYSILEKAPAAKDLFSFLANGVDEPT----NP

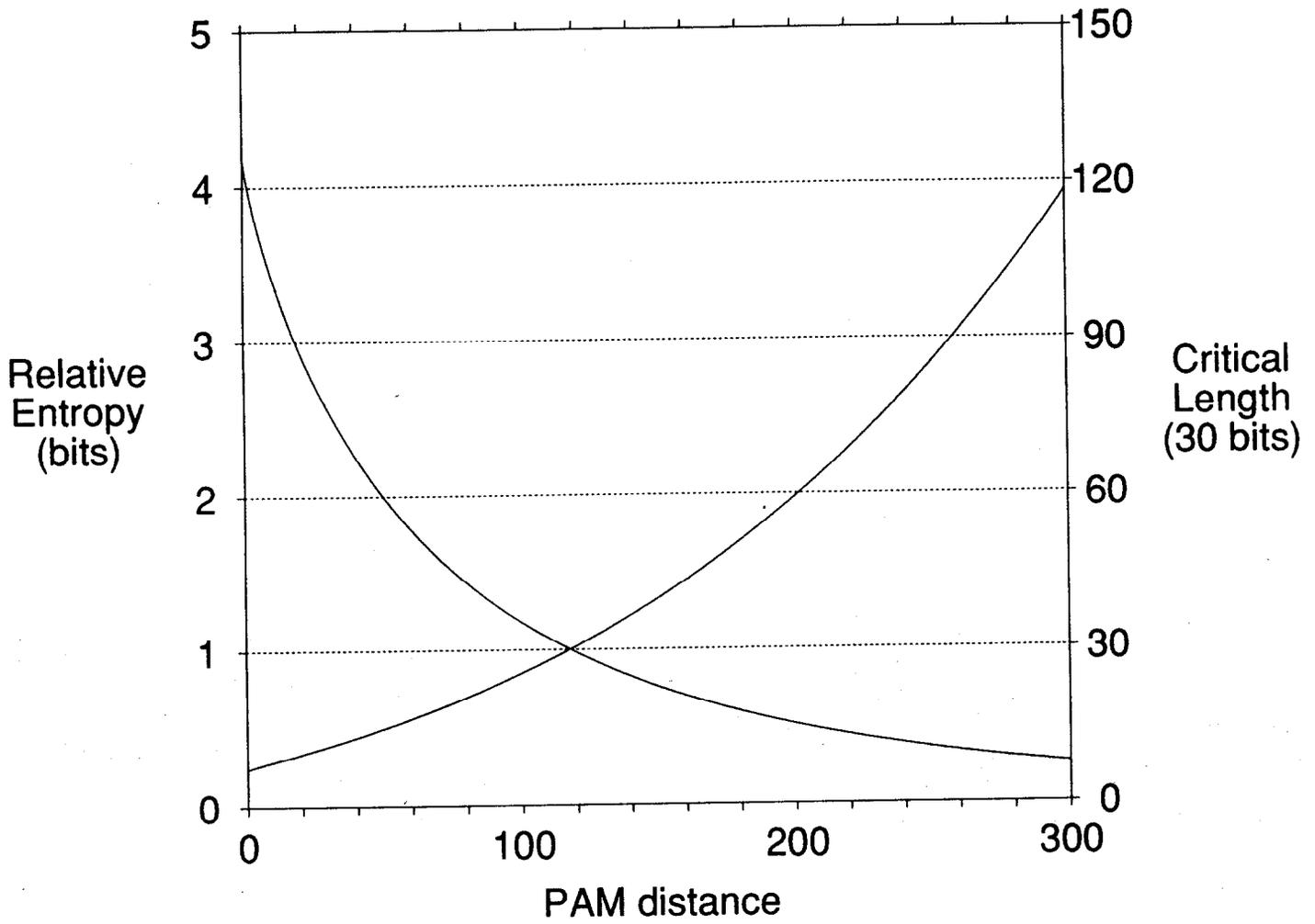
KVKAHGKKVLGAFSDGLAHLNLDKGTFA--TLSELHCDKLVHDPENFRLLGNVLVLCVLAHFGKEFTPPVQAAAYQKVVAGVANALAHKYH  
K H K D L A L H K DP F L G A A A  
KLTGHAEKLFALVRDSAGQLKASGTVVADAALGSVHAQKAVTDPQ-FVVVKEALLKTIKAAVGDKWSDELRAWEVAYDELAAAIKKA--

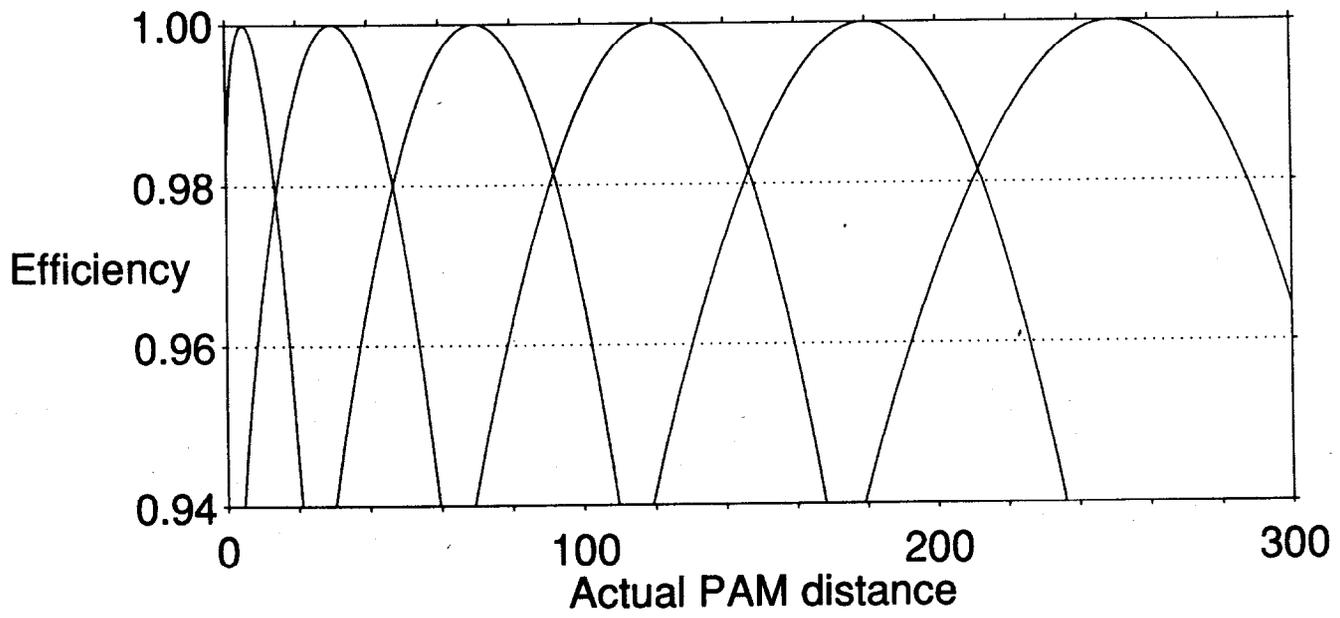
## RELATIVE ENTROPY

For alignments characterized by the amino acid pair frequencies  $q_{ij}$ , the maximum average score (information) achievable per alignment position is given by the formula

$$\sum_{i,j} q_{ij} \log \frac{q_{ij}}{p_i p_j}.$$

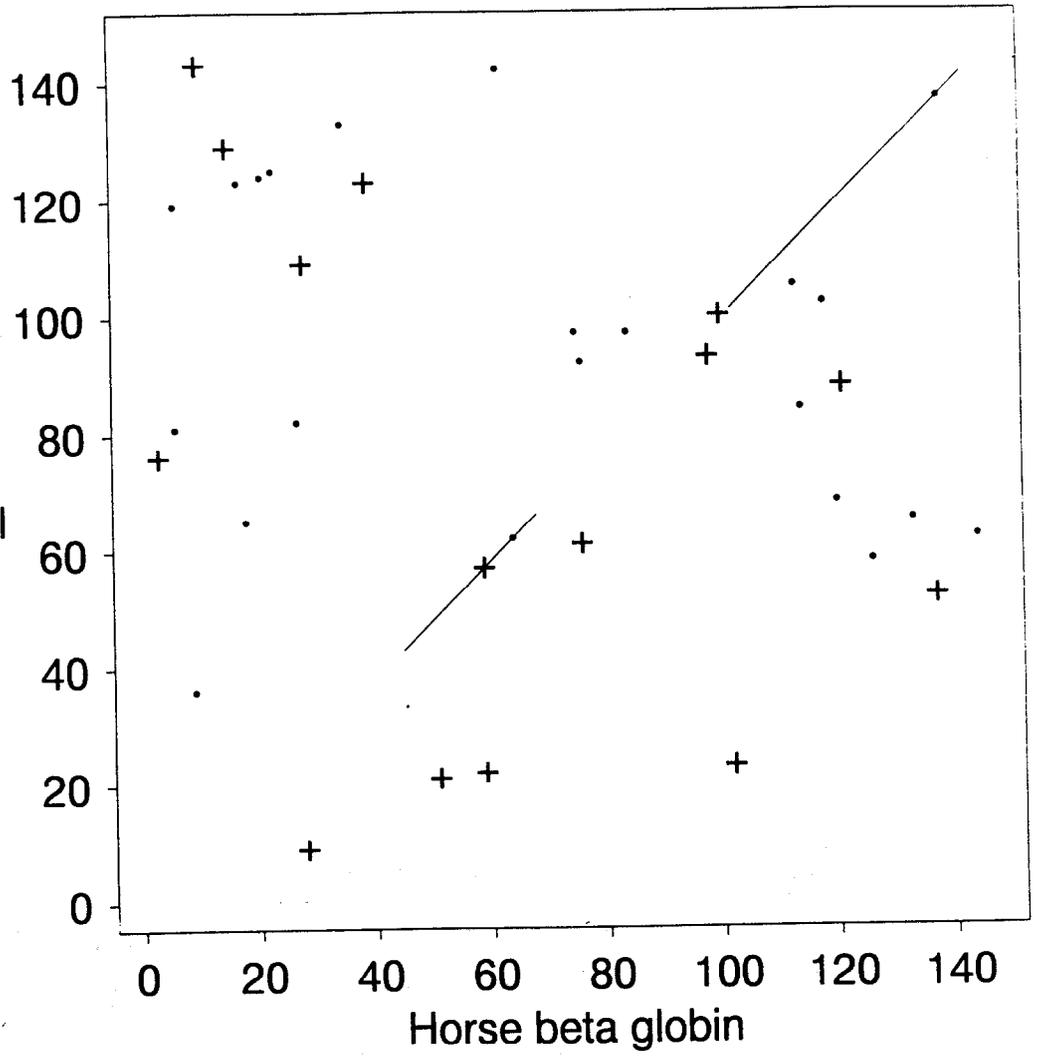
This average score can be attained only by using the appropriate matrix (JMB 219:555-565).



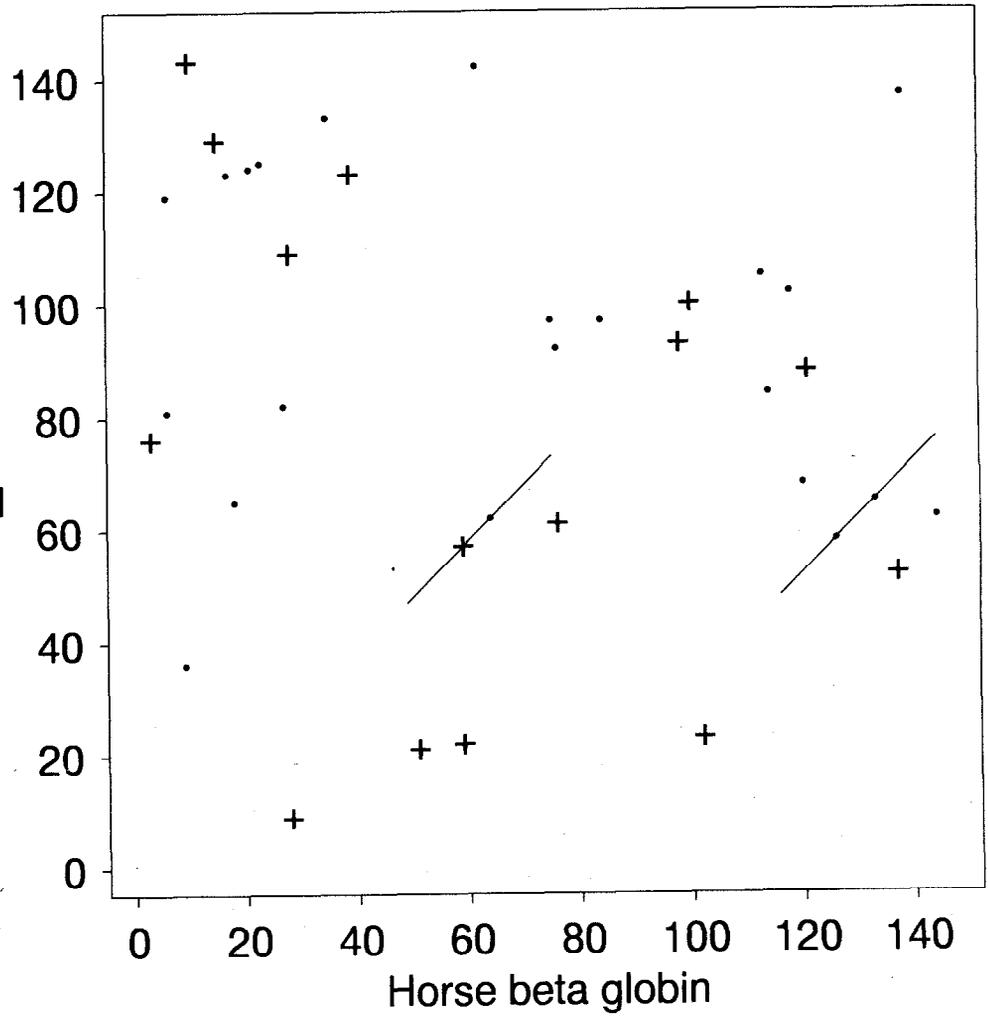




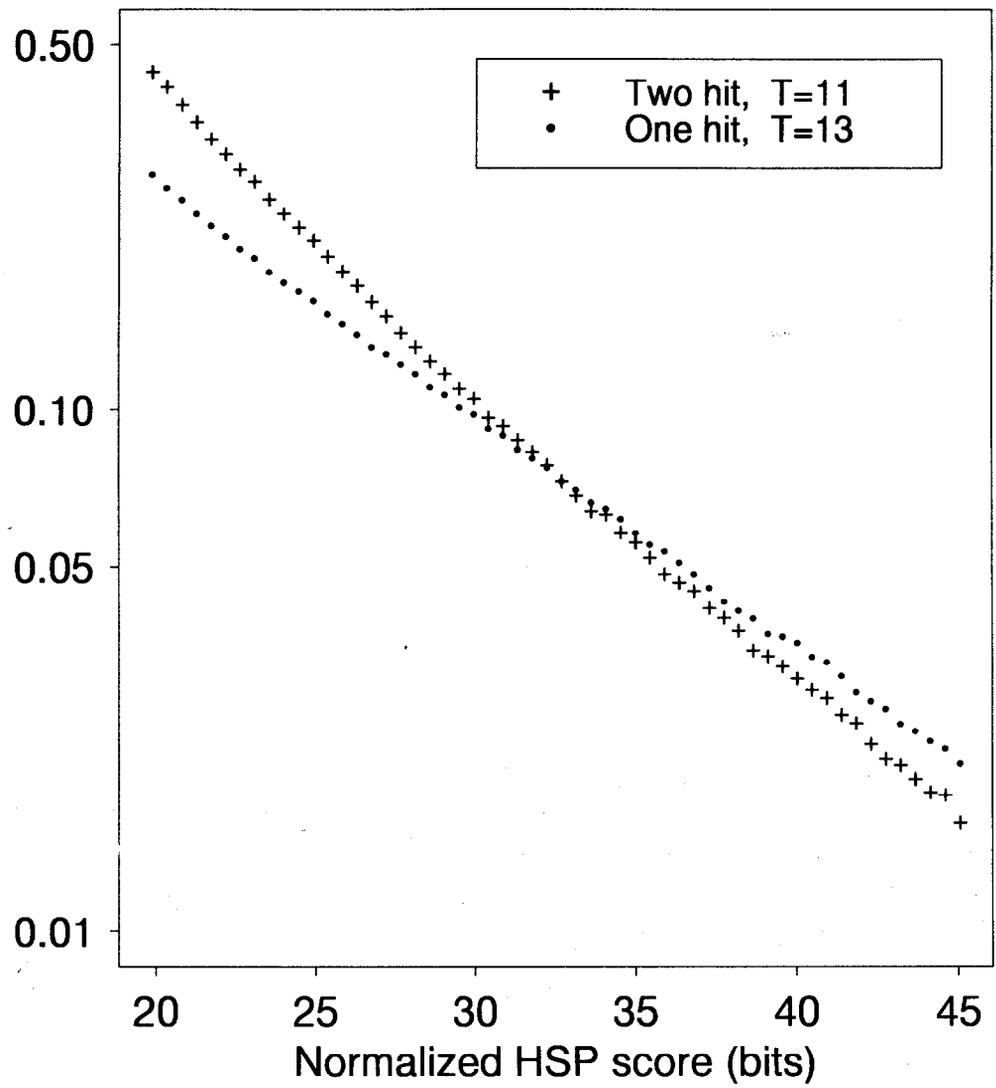
Broad bean  
leghemoglobin I



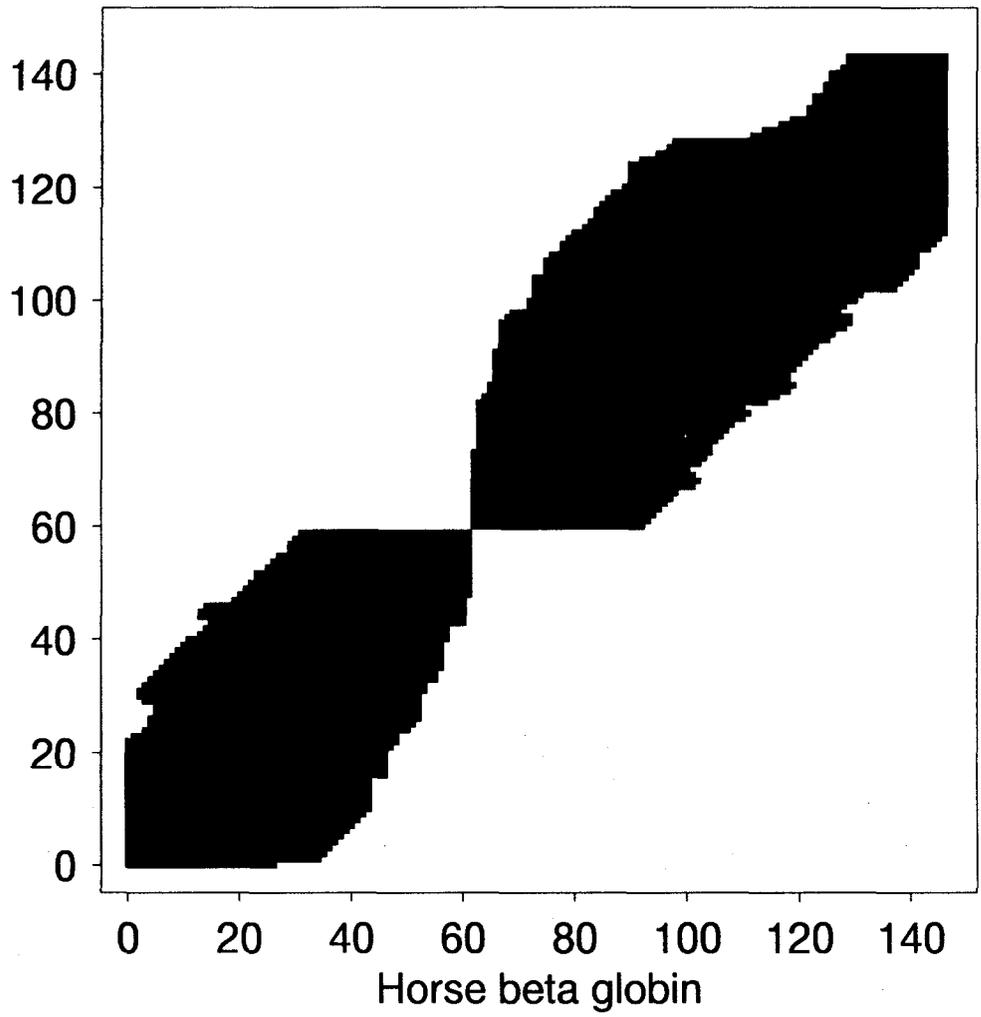
Broad bean  
leghemoglobin I



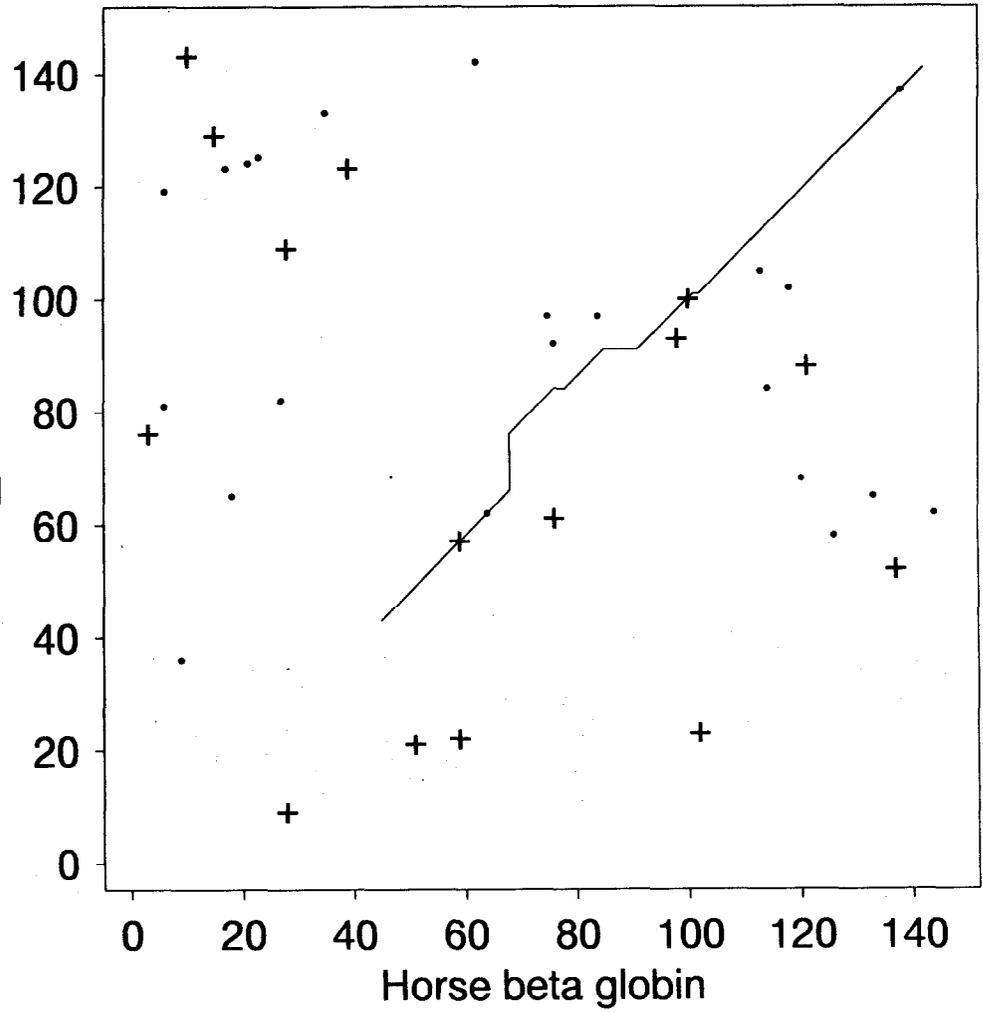
Probability of missing an HSP



Broad bean  
leghemoglobin I



Broad bean  
leghemoglobin I



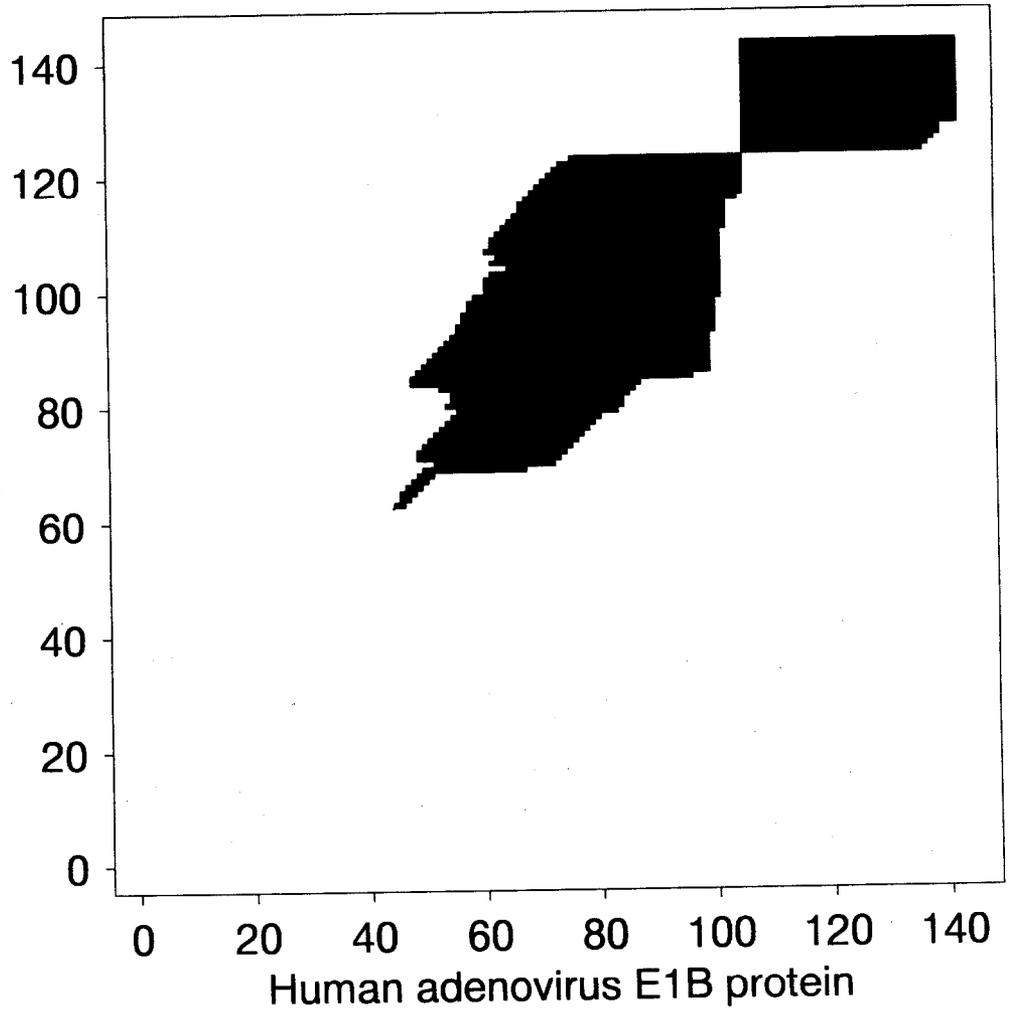
Leghemoglobin 43 FSFLKDSAGVVDSPKLGAAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS----- 90  
 L + V+ +PK+ AH +KV L + GE V LD G+

Beta globin 45 FGDLSNPGAVMGPNPKVKAHGKKV-----LHSFGEGVHHLNLIKGTFAALSE 90

Leghemoglobin 91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAWEVAYDGLATAI 140  
 +H K +DP +F ++ L+ + G ++ EL A+++ G+A A+

Beta globin 91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141

Broad bean  
leghemoglobin I



Relative times spent by the original and gapped BLAST programs on various algorithmic stages

	Overhead: database scanning, output, etc.	Calculating whether hits qualify for ungapped extension	Ungapped extensions	Gapped extensions
Original BLAST	8 ( 8%)		92 (92%)	
Gapped BLAST	8 (24%)	12 (37%)	5 (15%)	8 (24%)

(*Nucl. Acids Res.* **25**:3389-3402)

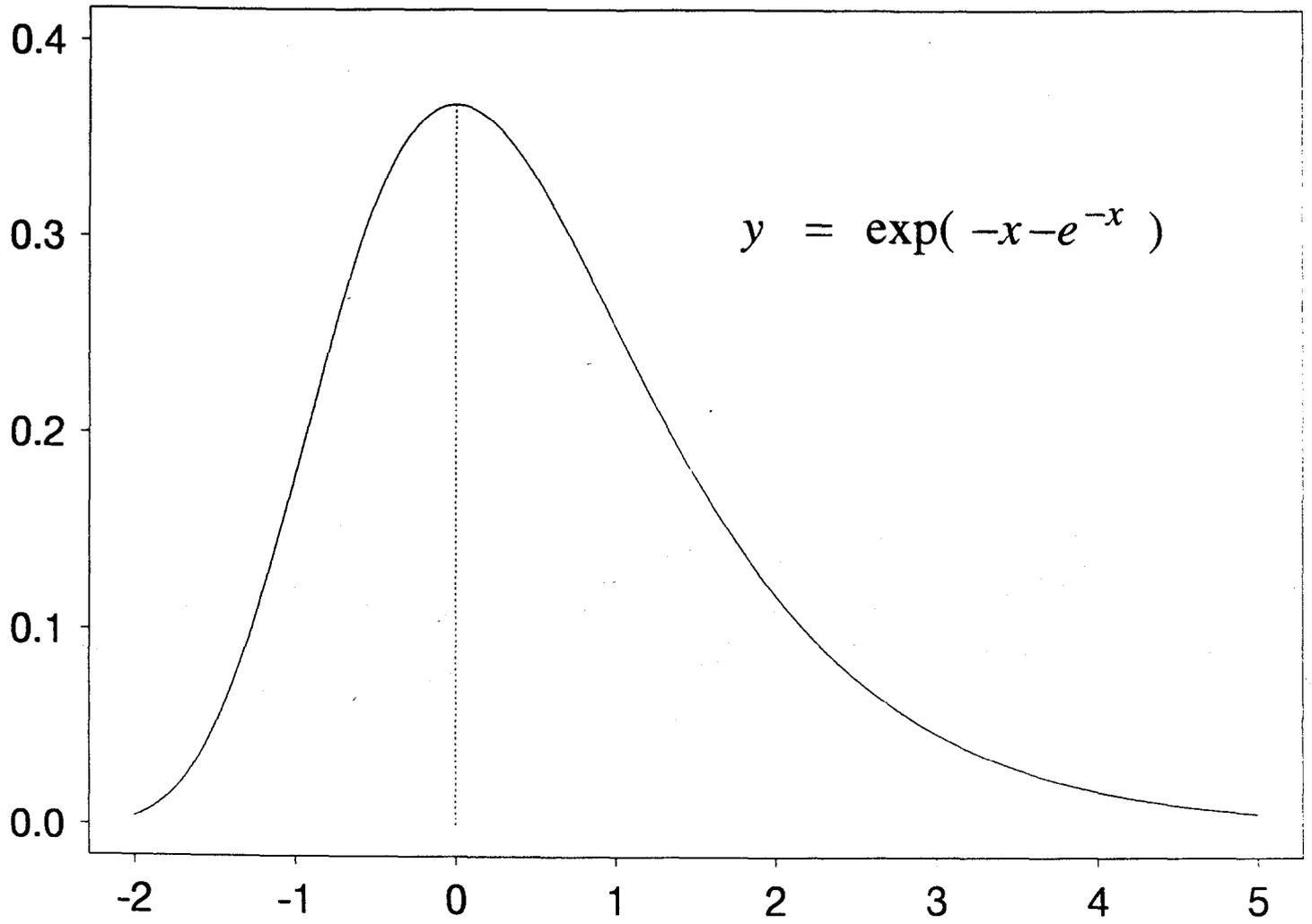
## Position-Specific Iterated BLAST (PSI-BLAST)

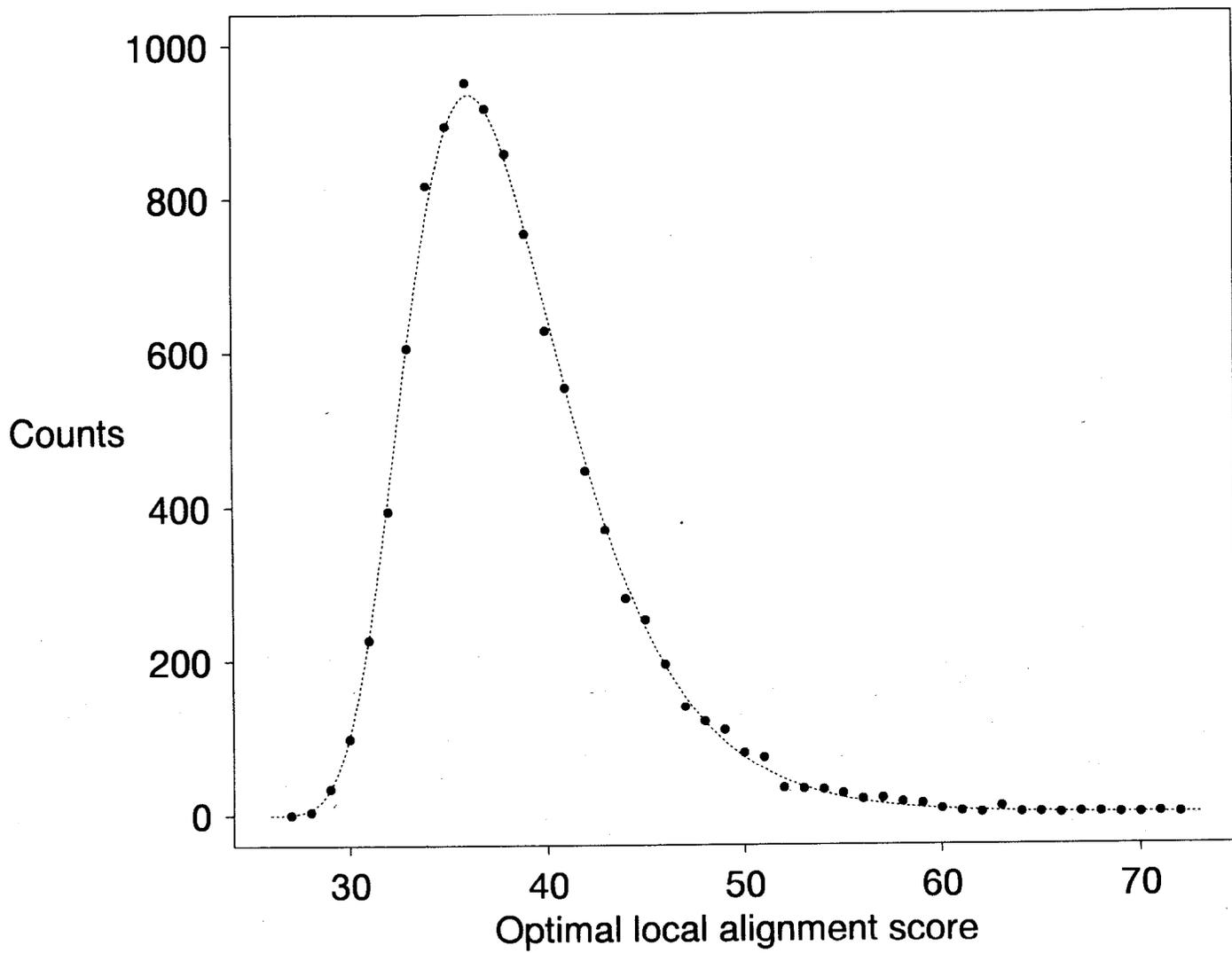
- a) Position specific score matrix the same length as query.
- b) Multiple alignment constructed from BLAST output.
- c) Sequences weighted on a column-by-column basis.
- d) Effective number of indep. observations estimated.
- e) Target freqs. derived using data-dependent pseudo-counts.
- f) Log-odds weight matrix scores calculated to scale.
- g) BLAST applied to position-specific score matrix.
- h) Statistical evaluation of results.
- i) Iteration.

Accession                      Alignment                      E-value

P49789		
P49779		8e-27
P49775		6e-18
Q11066		3e-07
Q09344		4e-05
P49378		0.001
P32084		0.002

# The Standard Extreme Value Distribution





Transformation of  $\lambda_u$  and  $K_u$  to  $\hat{\lambda}_g$  and  $\hat{K}_g$  when (10,1) affine gap costs are employed. The PSI-BLAST matrix is derived from 128 statistically significant local alignments from the comparison of influenza A virus hemagglutinin precursor to SWISS-PROT.

Scoring system	$\lambda_u$	$\hat{\lambda}_g$	$K_u$	$\hat{K}_g$
BLOSUM-62 matrix	0.3176	0.252	0.134	0.035
PSI-BLAST matrix	0.3175	0.254 (0.252)	0.154	0.040 (0.040)

The accuracy of PSI-BLAST statistics for the comparison of various query sequences with a shuffled version of SWISS-PROT.

Protein family	SWISS-PROT accession number of query	Low <i>E</i> -value	Number of seqs. with <i>E</i> -value	
			≤ 1	≤ 10
Serine protease	P00762	0.94	1	8
Serine protease inhibitor	P01008	1.5	0	9
Ras	P01111	1.1	0	9
Globin	P02232	8.2	0	2
Hemagglutinin	P03435	0.87	1	8
Interferon $\alpha$	P05013	0.11	2	11
Alcohol dehydrogenase	P07327	1.5	0	9
Histocompatibility antigen	P10318	0.0031	2	6
Cytochrome P450	P10635	0.46	1	15
Glutathione transferase	P14942	0.30	2	9
H <sup>+</sup> -transporting ATP synthase	P20705	0.79	2	10
Average (median or mean)		0.87	1.0	8.7

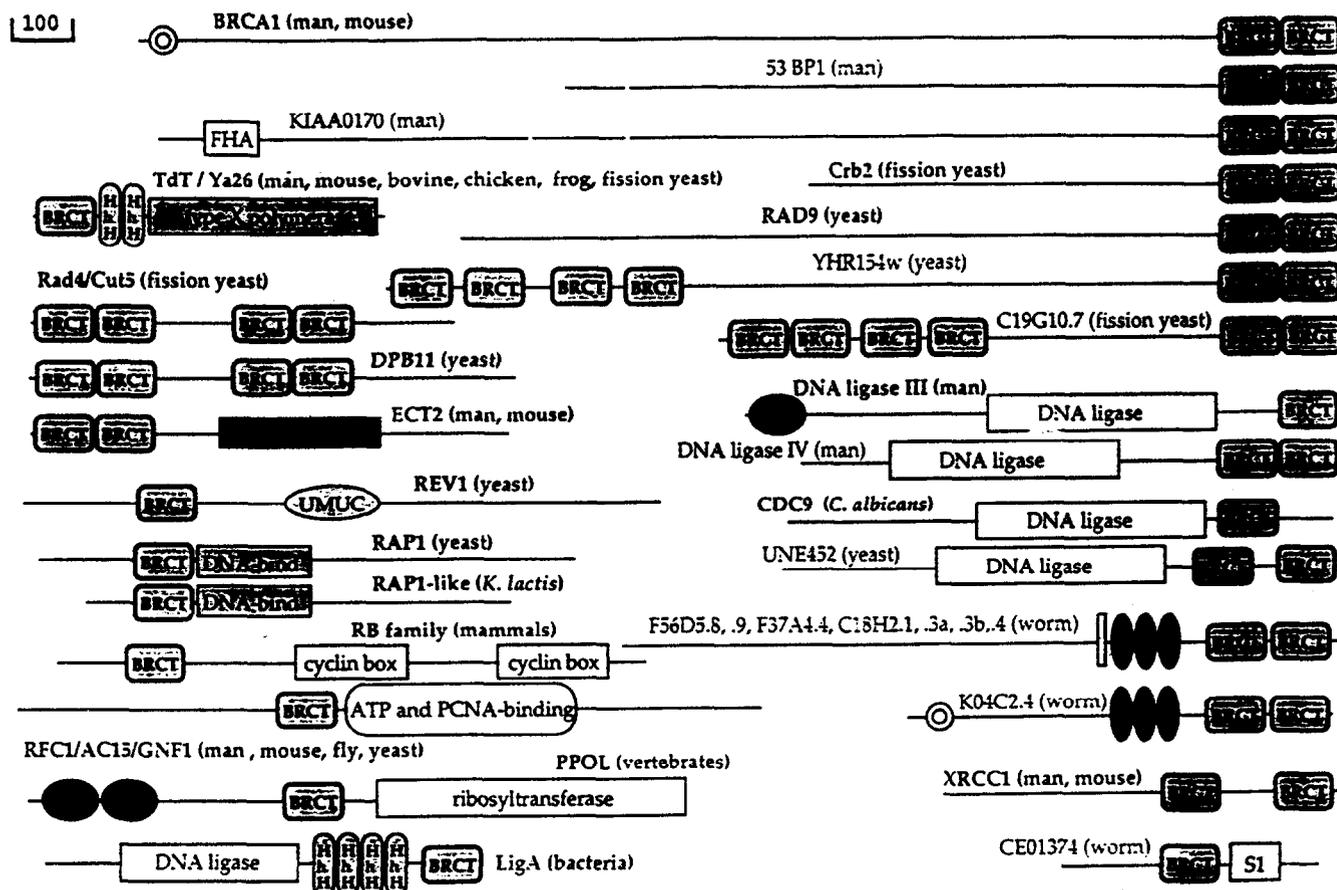
The number of SWISS-PROT sequences yielding alignments with  $E$ -value  $\leq 0.01$ , and relative running times, for Smith-Waterman and various versions of BLAST

Protein family	Query	Smith-Waterman	Original BLAST	Gapped BLAST	PSI-BLAST
Serine protease	P00762	275	273	275	286
Serine protease inhibitor	P01008	108	105	108	111
Ras	P01111	255	249	252	375
Globin	P02232	28	26	28	623
Hemagglutinin	P03435	128	114	128	130
Interferon $\alpha$	P05013	53	53	53	53
Alcohol dehydrogenase	P07327	138	128	137	160
Histocompatibility antigen	P10318	262	241	261	338
Cytochrome P450	P10635	211	197	211	224
Glutathione transferase	P14942	83	79	81	142
H <sup>+</sup> -transporting ATP synthase	P20705	198	191	197	207
Normalized running time		36	1.0	0.34	0.87

(*Nucl. Acids Res.* 25:3389-3402)

PSI-BLAST protein database search results using the C-terminus of BRCA1 as query

Protein	Species	GenBank ID number	PSI-BLAST iteration	E-value
BARD	<i>H. sapiens</i>	1710175	0	2e-06
T10M13.12 *	<i>A. thaliana</i>	2104545	1	4e-06
F26D2.b +	<i>C. elegans</i>	1914176	1	4e-04
KIAA0259 *	<i>H. sapiens</i>	1665785	1	0.001
F37D6.1	<i>C. elegans</i>	1418521	2	4e-06
C19G10.07	<i>S. pombe</i>	1723501	2	6e-05
KIAA0170	<i>H. sapiens</i>	1136400	2	0.002
53BP1	<i>H. sapiens</i>	488592	2	0.008
T13F2.3 *	<i>C. elegans</i>	1667334	3	2e-07
K04C2.4	<i>C. elegans</i>	470351	3	3e-07
T19E10.1	<i>C. elegans</i>	1067065	4	7e-04
Rad4/Cut5	<i>S. pombe</i>	730470	4	0.002
REV1	<i>S. cerevisiae</i>	132409	4	0.003
ECT2	<i>M. musculus</i>	423597	5	1e-04
XRCC1	<i>M. musculus</i>	627867	5	6e-04
Crb2	<i>S. pombe</i>	1449177	5	0.002
RAP1	<i>S. cerevisiae</i>	173558	5	0.006
TcEST030 #	<i>T. cruzi</i>	1536857	6	0.001
DPB11	<i>S. cerevisiae</i>	1352999	6	0.001
L8543.18	<i>S. cerevisiae</i>	1078075	6	0.010
SPAC6G9.12 *	<i>S. pombe</i>	1644324	7	4e-04
YM8021.03	<i>S. cerevisiae</i>	1078533	7	0.005
YHR154w	<i>S. cerevisiae</i>	731729	7	0.008
C36A4.8 *	<i>C. elegans</i>	1657667	7	0.010
UNE452	<i>S. cerevisiae</i>	1151000	8	8e-04
DNA ligase IV	<i>H. sapiens</i>	1706482	8	0.008
CDC9	<i>C. albicans</i>	1706483	9	0.006
DNA ligase	<i>T. scotoductus</i>	1352293	10	0.010
GNF1	<i>D. melanogaster</i>	544404	11	0.004
mutT #	<i>M. jannaschii</i>	2129134	15	0.008
RAD9	<i>S. cerevisiae</i>	131817	7	0.74
RAP1 homolog	<i>K. lactis</i>	422087	9	0.21
ZK675.2	<i>C. elegans</i>	599712	13	3.5
D90904 *	<i>Synechocystis sp.</i>	1652299	15	0.17
TDT	<i>M. domestica</i>	2149634	15	0.46
YGR103w	<i>S. cerevisiae</i>	1723693	16	0.017
Pescadillo *	<i>H. sapiens</i>	2194203	16	0.017
PPOL	<i>S. peregrina</i>	1709741	16	0.060



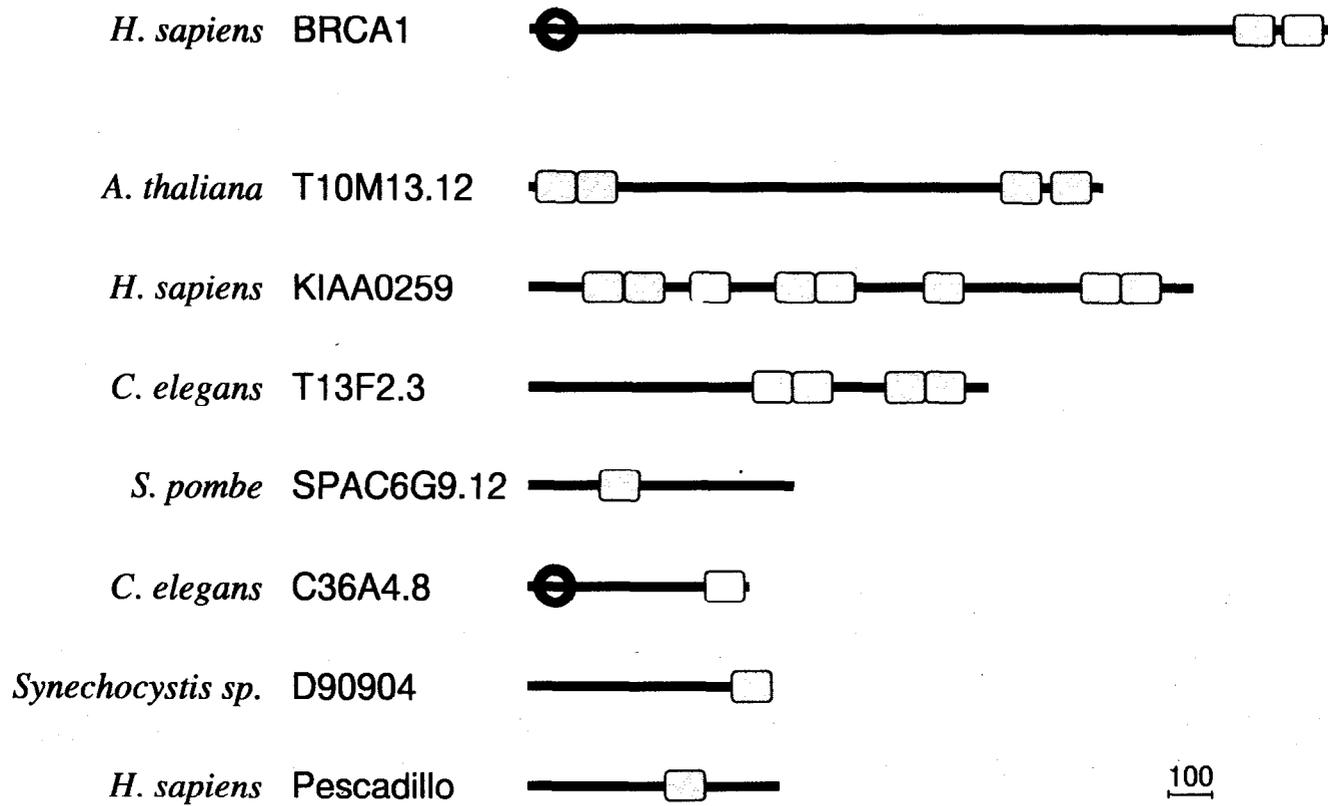
**Figure 2.** Domain organization of proteins containing BRCT domains. The proteins are shown roughly to scale as indicated by the bar in the upper left corner. The KIAA010 sequence is compressed as indicated by a broken line, and the 53BPI sequence is incomplete at the amino terminus. The names of proteins that have been functionally characterized are in bold type. In addition to the BRCT domain, other domains detected experimentally or by computer analysis are indicated. FHA is a putative nuclear signaling domain (23); AZF is a specific Zn finger domain found in PARP (designated PPOL in the figure) and DNA ligase III; HhH is the recently identified helix-hairpin-helix DNA binding domain (93); S1 is a putative RNA binding domain shared by bacterial ribosomal protein S1, polynucleotide phosphorylase, and yeast splicing factors (P. Bork, unpublished observations); RB has been reported to contain two cyclin box domains (94, 95), but the observed sequence similarity is very low; ANK indicates ankyrin repeats, and a double circle in BRCA1 and K04C2.4 indicates a RING finger. Only one representative for each set of proteins with similar modular architecture is included, e.g., only one of six worm paralogs that contain a transmembrane region (gray box) and ankyrin repeat. The species range is indicated for each domain architecture. Only one representative for each set of orthologs is included. Note that some of the proteins do not correspond to the annotation in the databases or to translations obtained by automatic procedures. For example, the yeast genes UNE407 and UNE452 were fused because UNE407 contains the amino-terminal portion of the DNA ligase domain and UNE452 contains the carboxy-terminal portion. Translation of *C. elegans* genes obtained by genomic sequencing was modified in order to optimize the alignment within the family. Specifically, C18H2.3 (PID: g474199) was split into two ORFs: in C18H2.4 (PID: g474200), additional putative exons were introduced.

tral region of PARP, which contains its single BRCT domain with significant similarity to the BRCT domain of RF, has been implicated in the protein's dimerization, but is not involved in DNA binding (43). The difference between the results obtained with RF-C and PARP requires further clarification, even though it may be a reflection of the actual diversity of the BRCT domain binding affinities.

### Modular architecture of BRCT domain-containing proteins

All members of the BRCT superfamily are large, multidomain proteins (Fig. 2). Many contain functionally characterized enzymatic domains, such as two unre-

lated types of DNA ligase, type X DNA polymerase (TdT), ADP-ribosyltransferase (PARP), and ATPase (RF-C). Other proteins in the superfamily contain additional common binding domains such as the RING finger in BRCA1 and an uncharacterized nematode protein, the DH domain in ECT2 and an uncharacterized yeast protein, the FHA domain in an uncharacterized human protein, the helix-hairpin-helix DNA binding domain in bacterial ligases and TdT, and ankyrin repeats in a family of uncharacterized nematode proteins. Yet other proteins contain highly conserved domains whose specific function is not known but are implicated in DNA repair, e.g., the UmuC domain in REV1. It is possible that some of these conserved domains possess yet uncharacterized enzymatic activities as demon-



**COLLABORATORS**  
**(Gapped BLAST and PSI-BLAST)**

Tom Madden, Jinghui Zhang, David Lipman (NCBI)

Alejandro Schaffer (NHGRI)

Zheng Zhang, Webb Miller (Penn. State)

*Nucl. Acids Res.* **25**:3389-3402.